A Few Variables Capture over 70% of State-Level Variance in U.S. Covid Deaths

Preprint, not peer reviewed

Joseph Sill, PhD
joe_sill@yahoo.com

June 18, 2021

Abstract

A U.S. state-level analysis of factors associated with covid-19 deaths reveals inequality (as defined by the Gini coefficient) to be far and away the strongest single-variable predictor, capturing 40% of variance in covid deaths and 49% of variance in all-cause excess deaths since the start of the pandemic. A linear regression model with 5 independent variables accounts for over 70% of variation in covid deaths, as does a 4-variable linear regression model for all-cause excess deaths. Similar models for covid and all-cause excess deaths since October 1 achieve similar results. Coefficients are highly significant ($p < 0.01$) in almost all cases. A consistent finding across all 4 models is that a state's relative humidity is strongly associated with fewer deaths after controlling for other factors. Lockdown stringency (as measured by the Oxford stringency index) is also strongly associated with fewer deaths for all models. Other significant factors for some models include population density, nursing home resident density, voting patterns shifting towards Donald Trump vs. prevous Republican candidates, and share of population under 18 years old. The models pass various robustness checks. The results are reproducible via an open-access data repository and Python notebook made available online.

1 Introduction

Covid-19 death rates vary widely across U.S. states, ranging from 297 per 100K population in New Jersey to below 50 per 100K in Alaska, Vermont and Hawaii as of 6/16/2021 [6]. The reasons for this variation remain hotly debated within the public sphere and many observers claim that the variation seems largely random. A recent article in a public affairs online magazine acknowledges some evidence for the effectiveness of lockdown measures but also observes that looking "at a list of states by their number of Covid-19 deaths per capita, it's hard to discern much of a pattern " [5]. A leading covid modeller, when looking at state-level data, said the data supports the theory that the virus is " inherently unpredictable " [12] and " the lack of correlations here suggest luck + randomness may play a large role" [13]. The 45th president of the United States recently issued a statement making the claim "blue state lockdowns didnt work " [20].

The effectivness of lockdown measures and (more generally) the reasons for US state-level covid death variation are of vital importance for a number of reasons. Helping answer these questions can inform policy decisions for the current, covid-19 pandemic as well as for future pandemics involving pathogens with similar characteristics. If the evidence were to show that lockdowns were indeed ineffective, their undesirable side effects could be avoided in the future. On the other hand, evidence of efficacy of lockdowns obviously strengthens the case for their continued implementation in the future. Additionally, if the perception among many that lockdowns were ineffective is widespread but is in fact incorrect, confidence in government and public health institutions may be undermined unnecessarily. Accurate assessments of cost-benefit tradeoffs of lockdown decisions can also help inform future elections, as voters attempt to evaluate the decisions of elected officials. Beyond simply looking at lockdowns, there is value in understanding which potential confounders are significant. Factors which may at first blush appear to be uncontrollable environmental characteristics (e.g. relative humidity) may in fact be subject to manipulation in some situations (e.g. indoors) which could help limit disease spread.

The public debates on the topic often lack rigor. For instance, Florida and California are commonly compared, but this level of analysis involves an N of 2 and typically little or no attempt to control for confounding variables aside from perhaps a casual appeal to similar temperatures. More rigorous statistical analysis is needed in order to inform such discusssions. Some

examples of such research already exist, such as [7]. This paper builds on such previous work. In addition to presenting analysis and results, another goal of the paper is to advertise a public data repository [17], which builds on a previously released repository presented at [10]. A Python notebook for reproducing the results in this paper is also available at [17].

The author acknowledges that modern academic standards for causal inference from data are high. No claim is being made that any of the results in this paper constitute rigorous proof of causal effects. Nonetheless, the explanatory variables have considerable causal plausibility and the statistical associations are highly significant after controlling for several confounders. It is possible that some or all of the associations presented here are not causal, but in that case, it is likely that knowledge of the dynamics of covid-19 spread will nonetheless be enhanced by explaining why these statistical associations exist.

Outline The remainder of this article is organized as follows. Section 2 desribes prior, related research. Data collection and modeling choices are described in Section 3. Regression results and robustness checks are presnted in Section 4. Possible reasons for the statistical associations are discussed in Section 5. Future research directions and broader lessons drawn from the work are in Section 6.

2 Previous work

State-level variation in covid deaths through Dec 1, 2020 are studied in [7]. Poverty and a few different measures of population density are found to be associated with more per-population covid deaths, while lockdown stringency is found to be negatively associated. Note that while the author of [7] also wrote an op-ed in the Orange County Register praising Florida's appreach as compared to California's [8], [7] does not include humidity as an explanatory variable. This paper builds on the findings in [7] in various ways. Inequality is presented as a variable which captures far more variance in covid deaths than poverty. Additionally, relative humidity and other explanatory variables not considered in [7] are found to be significant here. All-cause excess death models are presented as well as covid death models and various checks for robustness and spatial correlation of residuals are presented. More recent data is used.

More informally, a recent Twitter thread [11] looked at correlations between pairs of variables from the data source [10]. Substantial correlation between lockdown stringency and increases in unemployment rates was observed, while minimal correlation between covid death rates and any of the other variables examined was found. This analysis, however, did not include any multivariate regressions, nor did it look at metrics such as inequality, poverty, humidity, or changes in voting share for Trump vs. previous Republican presidential candidates. Furthermore, the District of Columbia was included in the correlation calculations even though DC has 9X the population density of any state.

The significance of humidity for airborne viral transmission has been noted in much research, for instance, [22] [15] and in a Twitter thread by a leading aerosol scientist [14]. Research indicates that increased humidity leads to larger , heavy droplets which fall out of the air sooner than in low-humidity conditions, as well as other influences humidity may have.

An association between inequality and covid mortality was noted early in the pandemic at the state leve in [21] and at the county level in [9].

Much other relevant research exists and a more thorough list of relevant citations will be provided in the journal-ready version of this work.

3 Methods

39 variables (listed in table 1) measuring various socieconomic, environmental and lockdown-policy traits of each of the 50 U.S. states (as well as DC) were gathered. 12 of the variables were taken from [10], a previously created repository. The remainder of the variables were collected from various publicly available sources, government and otherwise. The data is available for download in a spreadsheet at [17]. Although data for DC was collected, most of the analysis in the paper is performed with DC excluded. As [7] found, population density is a crucial variable for modeling covid deaths since the start of the pandemic and DC has 9X the population density of the densest state (New Jersey). Such an extreme outlier renders population density essentially useless in a regression with N=51.

The meaning of most of the variables should be either self-evident from the variable name or easily discernible from following the web links in table 1. Clarifications, if necessary, can be obtained by contacting the author. However, a few of the variables which are central to the results of the paper

are described here. The 'stringency_index' variable is the average value of the Oxford stringency index [18] since March 1,2020. The stringency index is a composite measure [4] which merges various policies such as bans on public gatherings, school closings, workplace closings, stay-at-home requirements and others into an overall one-number measure of lockdown stringency. The 'gini_inequality' variable is the Gini coefficient [1] of each state, obtained from [2], the 2019 edition of the American Community Survey. The Gini coefficient summarizes inequality in a mathematically graceful way by using the entire income distribution rather than arbitraty cutoffs such as top 1%, top 10%, bottom 10%, etc. Trump16McCain08Shift is the difference between the share of the presidential vote received by Donald Trump in 2016 and the share received by John McCain in 2008. This measure, however imperfect, is intended to detect enthusiasm in the state's populace for Donald Trump specifically rather than general Republican lean of the state. The results are nearly identical if Mitt Romney's share in 2012 is used rather than McCain's share in 2008. However, Romney's share of the vote in Utah in 2012 is an outlier- much more of an outlier than McCain's home-state share in Arizona. The results are therefore slightly more robust if McCain in 2008 is used as the baseline.

Variable	Source
afr_amer_pct	en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_African-American_population
cars_per_capita_pct	en.wikipedia.org/wiki/List_of_U.S._states_by_vehicles_per_capita
cig_use_pct	en.wikipedia.org/wiki/List_of_U.S._states_by_vehicles_per_capita
dem_margin_2020	Gu github
diabetes_pct	ccd.cdc.gov/Toolkit/DiabetesBurden/Prevalence
doctors_per_capita	beckershospitalreview.com/workforce/50-states-ranked-by-most-active-physicians-per-100-000-population.html
gini_inequality	www.census.gov/content/dam/Census/ library/publications/2020/acs/acsbr20-03.pdf
high_blood_pressure_pct	www.census.gov/content/dam/Census/library/publications/2020/acs/acsbr20-03.pdf
hispanic_latino_pct	en.wikipedia.org/wiki/List_of_U.S._states_by_Hispanic_and_Latino_population
income_per_capita	Gu github
mean_temperature	Gu github
median_age	Gu github
mex_amer_pct	en.wikipedia.org/wiki/List_of_U.S._states_by_Hispanic_and_Latino_population
McCain08Share	en.wikipedia.org/wiki/2008_United_States_presidential_election
multi_gen_household_pct	data.census.gov/cedsci/table?q=multigenerational&g=0100000US.04000.001&tid=ACSDT1Y2010.B11017
nursing_resid_per_pop	www.kff.org/other/state-indicator/number-of-nursing-facility-residents
obesity_rate	Gu github
over_65_pct	www.prb.org/resources/which-us-states-are-the-oldest/
pct_in_city_min_200k	en.wikipedia.org/wiki/List_of_United_States_cities_by_population
perc_25plus_with_bachelors	Gu github
perc_blue_collar_jobs	Gu github
perc_pop_at_least_1_dose	Gu github
perc_pop_nonwhite	Gu github
perc_urban	Gu github
population_per_sq_m	Gu github
poverty_rate	en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_poverty_rate
regular_church_pct	www.pewforum.org/religious-landscape-study/compare/attendance-at-religious-services/by/state/
relative_humidity	www.pewforum.org/religious-landscape-study/compare/attendance-at-religious-services/by/state/
res_per_household	www.indexmundi.com/facts/united-states/quick-facts/all-states/average-household-size
Romney12Share	en.wikipedia.org/wiki/2012_United_States_presidential_election
seldomornever_church_pct	www.pewforum.org/religious-landscape-study/compare/attendance-at-religious-services/by/state/
share_in_apts_pct	nmhc.org/research-insight/quick-facts-figures/quick-facts-resident-demographics/geography-of-apartment-residents/
stringency_index	raw.githubusercontent.com/OxCGRT/covid-policy-tracker/master/data/OxCGRT_latest.csv
Trump16McCain08Shift	-
Trump16Share	en.wikipedia.org/wiki/2016_United_States_presidential_election
under_18_pct	www.indexmundi.com/facts/united-states/quick-facts/all-states/percent-of-population-under-18
undocumented_pct	www.pewresearch.org/hispanic/interactives/u-s-unauthorized-immigrants-by-state/
uninsured_pct	www.beckershospitalreview.com/rankings-and-ratings/states-ranked-by-uninsured-rates.html
urb_index_538	fivethirtyeight.com/features/how-urban-or-rural-is-your-state-and-what-does-that-mean-for-the-2020-election/

Table 1: Variables and data sources

The high correlation of Gini-coefficient inequality with covid deaths and excess deaths is one simple result presented in the paper (note: Gini coefficient, gini inequality and the simple term 'inequality' are used interchangeably in this paper). However, the bulk of the paper focuses on multivariate regressions with a handful of variables. Many methods exists for variable selection for multivariate regression. Fully automated one-by-one variable selection such as forward or backward stepwise selection based on p-values has been criticized as invalidating the statistical significance tests on the coefficients [16]. Such procedures can also be path-dependent, i.e., variables can switch from insignificant to significant depending on which other variables have been included. Therefore, such fully-automated stepwise procedures based on p-values or adjusted R-squared were avoided. Alternatives include cross-validation [16] for confirming the predictive value of each variable and LASSO regression [19] [3], which provides a principled way to examine many

variables at the same time, discarding some as irrelevant and retaining others. Prior knowledge and strength of belief in the predictive value of the variables is also often advocated as a guiding rule for deciding which variables to include.

In this paper, the primary model variables recurring in many of the models presented (Gini coefficient, stringency index, relative humidity, population density, Trump-McCain shift and nursing home residents as a share of population) were arrived at as follows. The model in [7], where population density, stringency and poverty were found to be significant, was taken as a starting point. Data for additional variables was collected and an informal search for additional variables with highly significant p-values and which increased R-squared (as measured by both adjusted R-squared and cross-validation-estimated R-squared) were sought. In particular, the gini coefficient relationship was discovered by adding income_per_capita to a model with poverty and finding a positive association with income after controlling for poverty and other factors. Subsequently, for confirmation purposes, a lasso regression with a cross-validation-optimized penalization term was conducted on the full set of 39 variables and found to give very similar results in terms of variable selection as the previous criteria had yielded.

Standard errors which are robust to heteroscedasticity have become standard usage within many social science disciplines and therefore the regressions resuls presented here use robust standard errors. However, the results do not change much when using traditional standard errors. The notebook code supplied allows the user to obtain standard errors either way.

Although many public resources display results in terms of deaths per 100k population, the decision was made to present regression results here in terms of deaths per 330 million, i.e. deaths per total U.S. population. This mode of presentation makes the impact of the factors more relatable, since the big-picture headline U.S. death total of approximately 600K as of mid-June, 2021 is well known. Deaths per 330 million vs. per 100k is merely a matter of applying a different scale factor to the numbers, of course, and the choice does not change the substance of results in any way.

All independent variables are standardized to be have 0 mean and variance 1 (aka z-scored). The suffix '_stdized' present in the covariate names indicates this standardization.

4 Results

The main results of the paper concern multivariate regressions, but the strikingly strong statistical relationships between Gini coefficient and both covid deaths and all-cause excess deaths are worthy of contemplation before digging into the regressions results. Figure 1 shows a scatter plot of deaths/100K vs. Gini coefficient for the 50 U.S. states since the start of the pandemic. Not only is 40% R-squared a substantial amount of explained variance- it is also visually arresting to note how completely empty the lower right corner of the plot is (where high-inequality, low-covid-death states hypothetically would be). The relationship with all-cause excess death percentage, shown in Figure 2, is even stronger (49% R-squared).

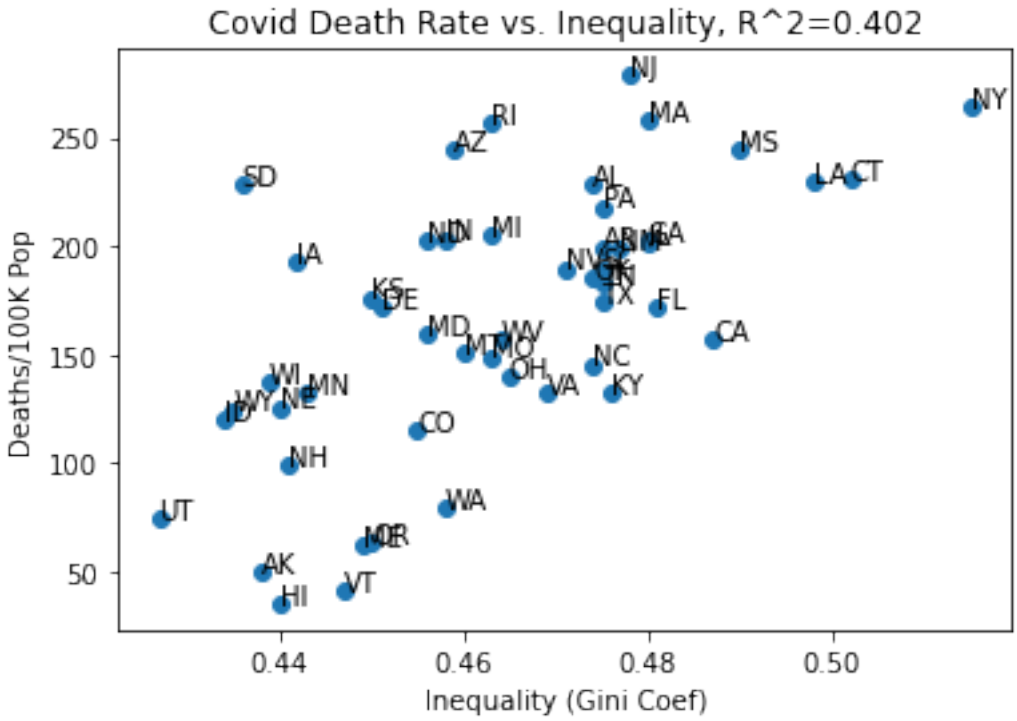


Figure 1: Covid death rate per 100K pop vs. Gini Coefficient

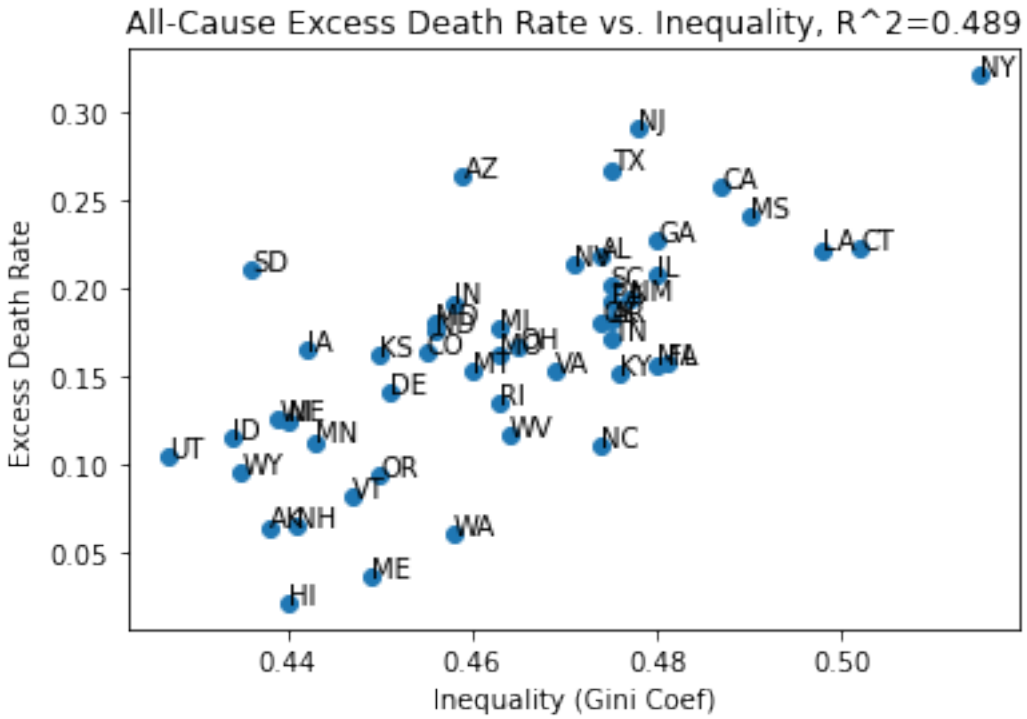
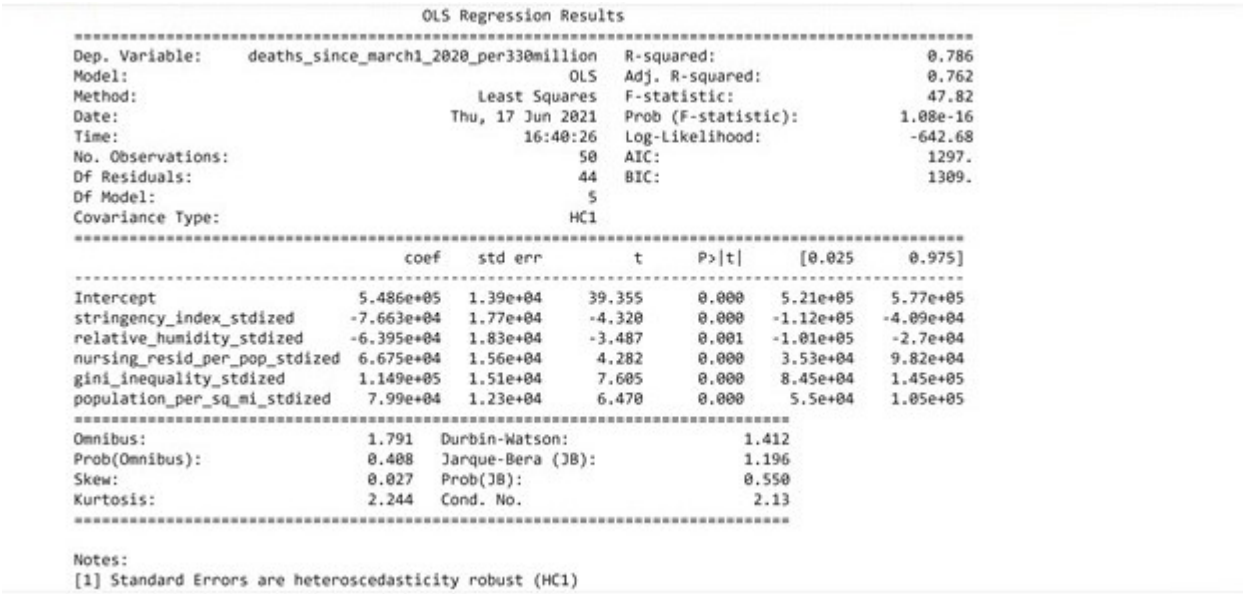


Figure 2: All-Cause Excess Death Rate vs. Gini Coefficient

Figure 3 shows regression results for covid deaths since March 1, 2020. All five factors are significant at the 99.9% level. An adjusted R-squared of 0.76 is achieved. 1 standard deviation in Gini coefficient is associated with 115k additional covid deaths nationally (per 330 million) after controlling for other factors. One standard deviation in population density is associated with 80k additional covid deaths. One standard deviation increase in lockdown stringency index is associated with 77k fewer covid deaths. One standard deviation increase in relative humidity is associated with 64k fewer deaths. One standard deviation increase in nursing home residents per population is associated with 67K additional deaths.



OLS Regression Results

Dep. Variable:	deaths_since_march1_2020_per330million	R-squared:	0.786
Model:	OLS	Adj. R-squared:	0.762
Method:	Least Squares	F-statistic:	47.82
Date:	Thu, 17 Jun 2021	Prob (F-statistic):	1.08e-16
Time:	16:40:26	Log-Likelihood:	-642.68
No. Observations:	50	AIC:	1297.
Df Residuals:	44	BIC:	1309.
Df Model:	5		
Covariance Type:	HC1		

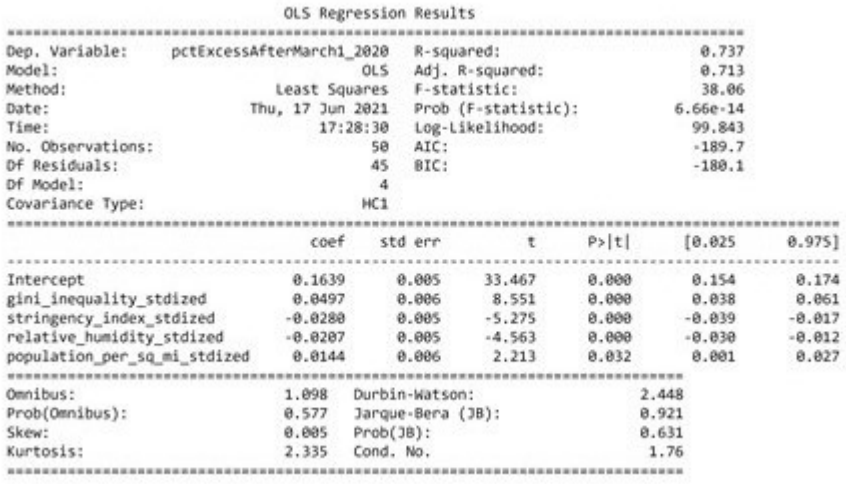
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.486e+05	1.39e+04	39.355	0.000	5.21e+05	5.77e+05
stringency_index_stdized	-7.663e+04	1.77e+04	-4.320	0.000	-1.12e+05	-4.09e+04
relative_humidity_stdized	-6.395e+04	1.83e+04	-3.487	0.001	-1.01e+05	-2.7e+04
nursing_resid_per_pop_stdized	6.675e+04	1.56e+04	4.282	0.000	3.53e+04	9.82e+04
gini_inequality_stdized	1.149e+05	1.51e+04	7.605	0.000	8.45e+04	1.45e+05
population_per_sq_mi_stdized	7.99e+04	1.23e+04	6.470	0.000	5.5e+04	1.05e+05

Omnibus:	1.791	Durbin-Watson:	1.412
Prob(Omnibus):	0.408	Jarque-Bera (JB):	1.196
Skew:	0.027	Prob(JB):	0.550
Kurtosis:	2.244	Cond. No.	2.13

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)

Figure 3: Regression results for Covid deaths since 3/1/2020

Figure 4 shows regression results for all-cause excess death percentage since March 1, 2020. One standard deviation increase in Gini coefficient is associated with a 5% increase in all-cause excess death percentage nationally. One standard deviation increase in lockdown stringency is associated with a 2.8% decrease nationally. One standard deviation increase in relative humidity is associated with a 2.1% decrease nationally. All 3 factors are significant at the 99.9% level. One standard deviation increase in population density is associated with a 1.4% increase in all-cause excess death percentage nationally. Population density is only significant at the 95% level. Population density was included because it decreases the spatial correlation of the residuals to a minimal level. It is unclear why nursing home population does not show up as significant here, but one possibility is that a portion of nursing home deaths were deaths of people who would otherwise have lived less than 15 additional months. Such deaths would not impact all-cause excess deaths measured in totality since March 1, 2020.



OLS Regression Results

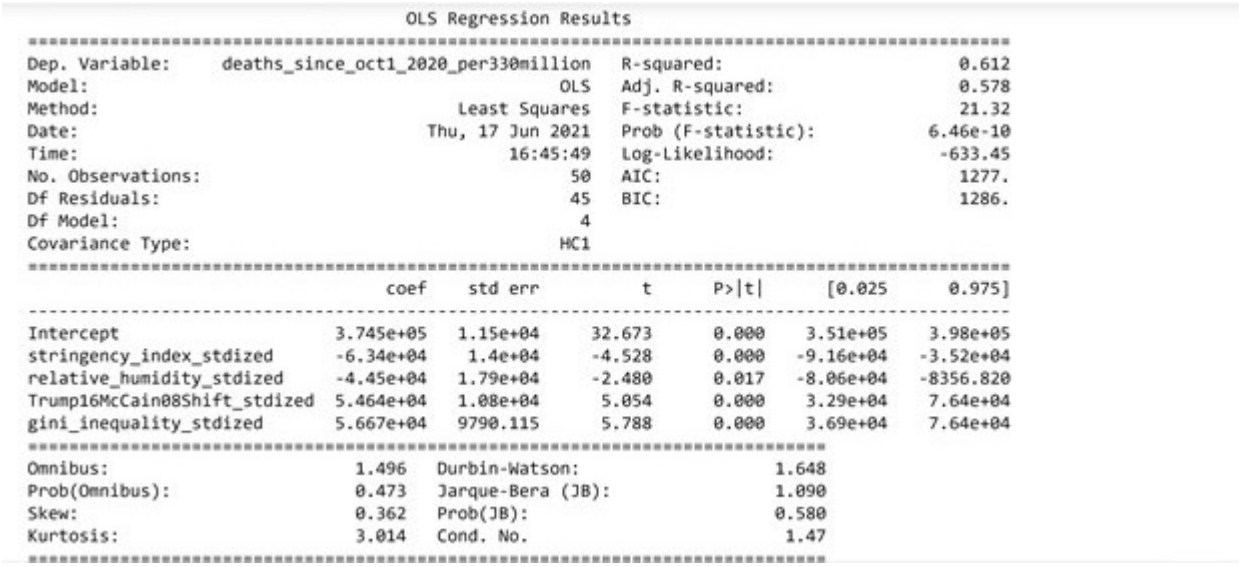
Dep. Variable:	pctExcessAfterMarch1_2020	R-squared:	0.737
Model:	OLS	Adj. R-squared:	0.713
Method:	Least Squares	F-statistic:	38.06
Date:	Thu, 17 Jun 2021	Prob (F-statistic):	6.66e-14
Time:	17:28:30	Log-Likelihood:	99.843
No. Observations:	50	AIC:	-189.7
Df Residuals:	45	BIC:	-180.1
Df Model:	4		
Covariance Type:	HC1		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1639	0.005	33.467	0.000	0.154	0.174
gini_inequality_stdized	0.0497	0.006	8.551	0.000	0.038	0.061
stringency_index_stdized	-0.0280	0.005	-5.275	0.000	-0.039	-0.017
relative_humidity_stdized	-0.0207	0.005	-4.563	0.000	-0.030	-0.012
population_per_sq_mi_stdized	0.0144	0.006	2.213	0.032	0.001	0.027

Omnibus:	1.098	Durbin-Watson:	2.448
Prob(Omnibus):	0.577	Jarque-Bera (JB):	0.921
Skew:	0.005	Prob(JB):	0.631
Kurtosis:	2.335	Cond. No.	1.76

Figure 4: Regression results for all-cause excess deaths since 3/1/2020

Figure 5 shows regression results for covid deaths since October 1. The same stringency, humidity and Gini coefficient factors which appear in the covid death model since the start of the pandemic show up again here. Another significant factor is the difference between the share of vote received by Donald Trump in the 2016 presidential election vs. the share of vote for John McCain in 2008. The interpretation of this variable will be examined more in the discussion section, but one interpretation is as a measure of enthusiasm for Donald Trump specifically rather than for the Republican party generally.



OLS Regression Results

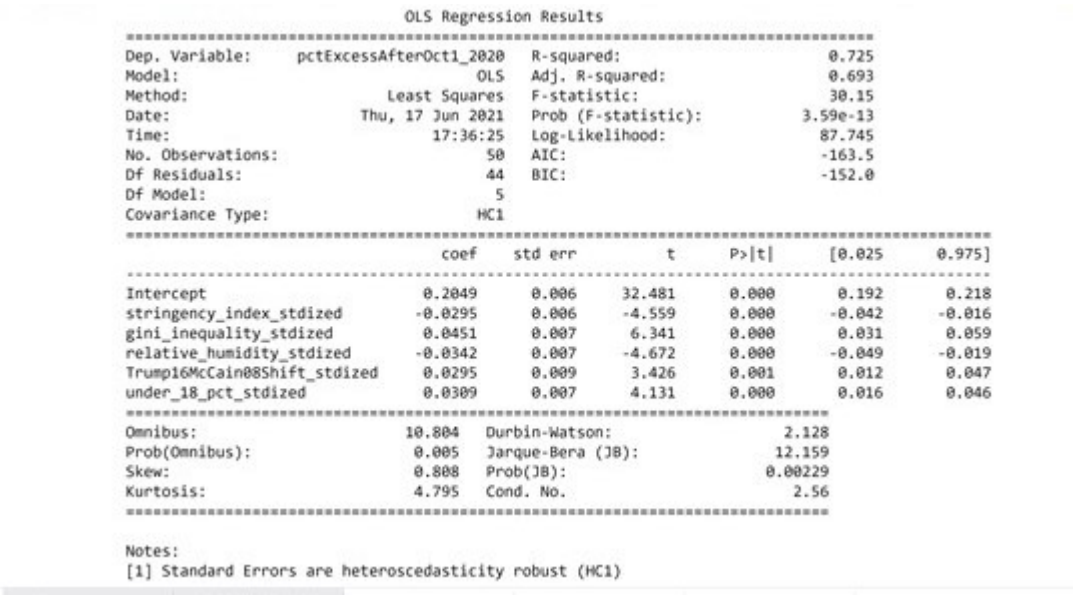
Dep. Variable:	deaths_since_oct1_2020_per330million	R-squared:	0.612
Model:	OLS	Adj. R-squared:	0.578
Method:	Least Squares	F-statistic:	21.32
Date:	Thu, 17 Jun 2021	Prob (F-statistic):	6.46e-10
Time:	16:45:49	Log-Likelihood:	-633.45
No. Observations:	50	AIC:	1277.
Df Residuals:	45	BIC:	1286.
Df Model:	4		
Covariance Type:	HC1		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.745e+05	1.15e+04	32.673	0.000	3.51e+05	3.98e+05
stringency_index_stdized	-6.34e+04	1.4e+04	-4.528	0.000	-9.16e+04	-3.52e+04
relative_humidity_stdized	-4.45e+04	1.79e+04	-2.480	0.017	-8.06e+04	-8356.820
Trump16McCain08Shift_stdized	5.464e+04	1.08e+04	5.054	0.000	3.29e+04	7.64e+04
gini_inequality_stdized	5.667e+04	9790.115	5.788	0.000	3.69e+04	7.64e+04

Omnibus:	1.496	Durbin-Watson:	1.648
Prob(Omnibus):	0.473	Jarque-Bera (JB):	1.090
Skew:	0.362	Prob(JB):	0.580
Kurtosis:	3.014	Cond. No.	1.47

Figure 5: Regression results for Covid deaths since 10/1/2020

Figure 6 shows regression results for all-cause excess deaths since October 1. In addition to the 4 variables included for covid deaths since October 1, one standard deviation increase in the share of the population under 18 is associated with a 3% increase in excess death percentage. Unlike the other factors, no obvious causal mechanism for the under-18-share factor is apparent to the author of the paper, but it shows up as highly significant. The association may be due to some as-yet-undetermined confounder, but in any case, the association is worth further study.



OLS Regression Results

Dep. Variable:	pctExcessAfterOct1_2020	R-squared:	0.725
Model:	OLS	Adj. R-squared:	0.693
Method:	Least Squares	F-statistic:	30.15
Date:	Thu, 17 Jun 2021	Prob (F-statistic):	3.59e-13
Time:	17:36:25	Log-Likelihood:	87.745
No. Observations:	50	AIC:	-163.5
Df Residuals:	44	BIC:	-152.0
Df Model:	5		
Covariance Type:	HC1		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.2049	0.006	32.481	0.000	0.192	0.218
stringency_index_stdized	-0.0295	0.006	-4.559	0.000	-0.042	-0.016
gini_inequality_stdized	0.0451	0.007	6.341	0.000	0.031	0.059
relative_humidity_stdized	-0.0342	0.007	-4.672	0.000	-0.049	-0.019
Trump16McCain08Shift_stdized	0.0295	0.009	3.426	0.001	0.012	0.047
under_18_pct_stdized	0.0309	0.007	4.131	0.000	0.016	0.046

Omnibus:	10.804	Durbin-Watson:	2.128
Prob(Omnibus):	0.005	Jarque-Bera (JB):	12.159
Skew:	0.808	Prob(JB):	0.00229
Kurtosis:	4.795	Cond. No.	2.56

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)

Figure 6: Regression results for all-cause excess deaths since 10/1/2020

4.1 Robustness Checks

Any regression conducted on data points with spatial relationships should check for spatial correlation of the residuals. Correlation of residuals invalidates the statistical significance tests of ordinary least squares, which relies on an assumption of i.i.d gaussian residuals. To check for spatial correlation, the correlation of each state's residual with the average residual of the states which border that state was computed and is shown in 4.1. Residual correlation with bordering states is minimal and is in fact slightly negative for 3 of the 4 models. The slight negative correlation is probably only small-sample noise but nonetheless strengthens evidence that positive spatial correlation is not a concern here. More formal spatial correlation tests may be conducted in future versions of this work. Figure 7 shows a color-coded representation of residuals for the covid-deaths-since-March1-2020 model, with blue indicating large negative residuals, red indicating large positive and purple indicating small magnitude residuals. No spatial pattern is evident. Analogous maps for the 3 other models are omitted for space consideration reasons but can be obtained by running the notebook code.

Model	Corr of Resid with Avg Border State Resid
covid deaths since March 1	-0.04
excess death pct since March 1	0.15
covid deaths since Oct 1	-0.16
excess deaths since Oct 1	-0.22

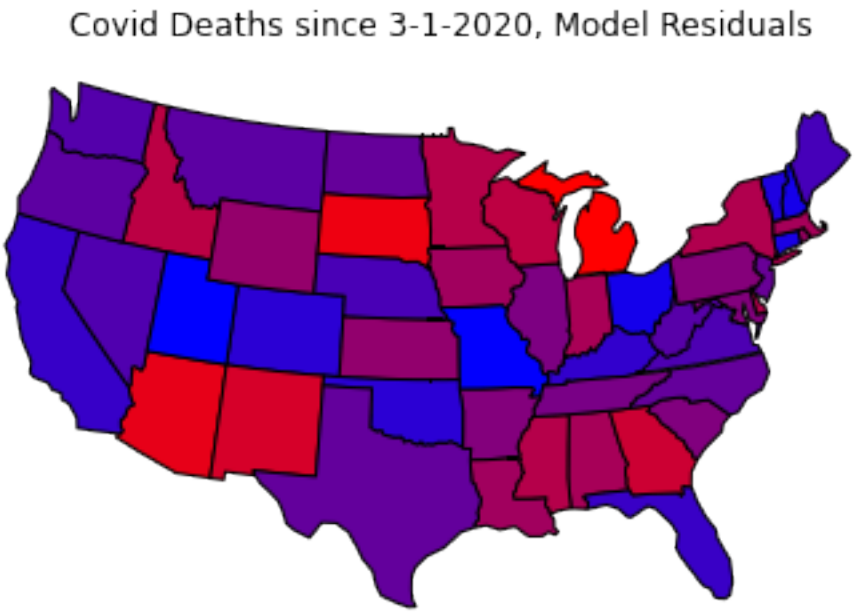


Figure 7: Model Resids, blue implies large neg, red implies large pos

When dealing with small sample sizes, additional confidence in regression results can be obtained by confirming that the results are not dependent on any individual data point. Accordingly, each of the 4 regressions were re-run with each of the 50 states removed and the highest p_value for any state removal was recorded for each coefficient. Tables 2, 3, 4, 5 display the results. In almost all cases, the coefficients remain highly significant with the removal of any state. The one major exception is that the significance of population density for all-cause excess since 3/1/2020 is not robust to the removal of New Jersey. While the intuitive case for the significance of population density for covid spread is clear, a case could be made that it cannot be distinguished

with high confidence from a story in which the dense northeastern states simply had the bad luck of getting hit hard with covid early on, particularly given that density does not show up in the since-October models. While humidity is only significant at the 90% level for covid deaths since October 1 after removing Arizona, the overall case for the significance of humidity seems strong, given that it remains highly significant with the removal of any state for the other 3 models.

Variable	**Max PValue**	**State Removed**
stringency_index_stdized	0.001	HI
relative_humidity_stdized	0.013	NM
nursing_resid_per_pop_stdized	0.000	AL
gini_inequality_stdized	0.000	AL
population_per_sq_mi_stdized	0.000	AL

Table 2: Max P_values with any state removed, covid deaths since 3/1/2020

Variable	**Max PValue**	**State Removed**
stringency_index_stdized	0.000	AL
relative_humidity_stdized	0.000	NM
nursing_resid_per_pop_stdized	0.000	AL
gini_inequality_stdized	0.000	AL
population_per_sq_mi_stdized	0.199	NJ

Table 3: Max P_values with any state removed, excess death pct since 3/1/2020

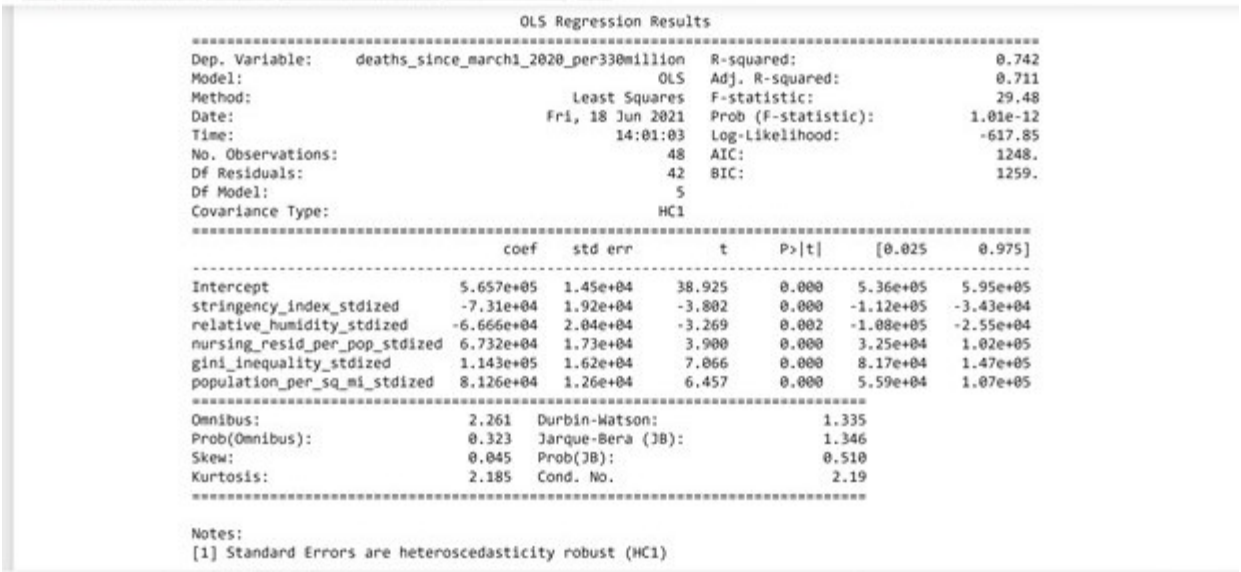
Variable	**Max PValue**	**State Removed**
stringency_index_stdized	0.001	HI
relative_humidity_stdized	0.074	AZ
Trump16McCain08Shift_stdized	0.000	AL
gini_inequality_stdized	0.000	AL

Table 4: Max P_values with any state removed, covid deaths since 10/1/2020

Variable	Max PValue	State Removed
stringency_index_stdized	0.001	HI
relative_humidity_stdized	0.001	AZ
Trump16McCain08Shift_stdized	0.015	UT
gini_inequality_stdized	0.000	AL
under_18_pct_stdized	0.003	NH

Table 5: Max Pvalues with any state removed, excess deaths since 10/1/2020

Alaska and Hawaii are both very high humidity, low covid death states and are arguably also outliers from the point of view of the difficulty of travel to and from those 2 states, so it is worth verifying whether the results are robust to removal of those 2 states. Figure 8 shows regression results with both states removed. The coefficients are nearly identical to those with the states included.



OLS Regression Results

Dep. Variable:	deaths_since_march1_2020_per330million	R-squared:	0.742
Model:	OLS	Adj. R-squared:	0.711
Method:	Least Squares	F-statistic:	29.48
Date:	Fri, 18 Jun 2021	Prob (F-statistic):	1.01e-12
Time:	14:01:03	Log-Likelihood:	-617.85
No. Observations:	48	AIC:	1248.
Df Residuals:	42	BIC:	1259.
Df Model:	5		
Covariance Type:	HC1		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.657e+05	1.45e+04	38.925	0.000	5.36e+05	5.95e+05
stringency_index_stdized	-7.31e+04	1.92e+04	-3.802	0.000	-1.12e+05	-3.43e+04
relative_humidity_stdized	-6.666e+04	2.04e+04	-3.269	0.002	-1.08e+05	-2.55e+04
nursing_resid_per_pop_stdized	6.732e+04	1.73e+04	3.900	0.000	3.25e+04	1.02e+05
gini_inequality_stdized	1.143e+05	1.62e+04	7.066	0.000	8.17e+04	1.47e+05
population_per_sq_mi_stdized	8.126e+04	1.26e+04	6.457	0.000	5.59e+04	1.07e+05

Omnibus:	2.261	Durbin-Watson:	1.335
Prob(Omnibus):	0.323	Jarque-Bera (JB):	1.346
Skew:	0.045	Prob(JB):	0.510
Kurtosis:	2.185	Cond. No.	2.19

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)

Figure 8: regression results for covid deaths since 3/1/2020 with Hawaii and Alaska removed.

To confirm that each variable adds out-of-sample predictive value to each model, 10-fold cross validated r-squareds were computed after removing each variable for each of the 4 models. Results are shown in tables 6, 7, 8 and 9. All

variables add substantial predictive accuracy to all models, with the possible exception of density for excess deaths since 3/1/2020, which was included for the sake of reducing spatial correlation, as previously mentioned.

None	0.708
stringency_index_stdized	0.602
relative_humidity_stdized removed	0.620
nursing_resid_per_pop_stdized	0.612
gini_inequality_stdized	0.380
population_per_sq_mi_stdized	0.592

Table 6: Cross-validated r_squared with variable removed, covid deaths since 3/1/2020

None	0.653
stringency_index_stdized	0.470
relative_humidity_stdized	0.548
gini_inequality_stdized	0.074
population_per_sq_mi_stdized	0.641

Table 7: Cross-validated r-squared with variable removed, excess death pct since 3/1/2020

None	0.474
stringency_index_stdized	0.237
relative_humidity_stdized removed	0.389
gini_inequality_stdized	0.255
Trump16McCain08Shift_stdized	0.292

Table 8: Cross-validated r-squared with variable removed, covid deaths since 10/1/2020

None	0.602
stringency_index_stdized	0.522
relative_humidity_stdized removed	0.426
gini_inequality_stdized	0.216
Trump16McCain08Shift_stdized	0.497
under_18_pct_stdized	0.546

Table 9: Cross-validated r_squared with variable removed, excess deaths since 10/1/2020

Tables 10, 11, 12, 13 show the coefficients obtained from a lasso regression with cross-validation-optimized penalization parameter. Any of the 39 variables not included in these tables was zeroed-out by the lasso regression. The significances of gini coefficient, lockdown stringency and humidity are confirmed by their consistent appearance among the largest-magnitude coefficients across all 4 models. The one exception is that lasso for covid deaths since October 1 assigns a much bigger coefficient to poverty than gini coefficient. It would be a defensible choice to include poverty rather than gini for this model, but these 2 variables are highly correlated and the overall themes of the paper would not change much if poverty were swapped in for gini coefficient for that one model. Racial and ethnic population percentages do also show up with substantial coefficients in the lasso regressions. The absence of these variables in the regression models presented in the paper should not be taken as denying that racism (structural or otherwise) has played a significant role in covid death trends in the United States. Alternative models which include some of these factors would certainly be defensible. Structural racism and inequality are intertwined in the United States to a degree that they are difficult to disentangle statistically with a dataset where N=50. The regression models in this paper were chosen with parsimony as an important criterion.

Variable	**Lasso coef**
population_per_sq_mi_stdized	83956
gini_inequality_stdized	61856
stringency_index_stdized	-59555
nursing_resid_per_pop_stdized	55570
relative_humidity_stdized	-47139
poverty_rate_stdized	37433
afr_amer_pct_stdized	26310
hispanic_latino_pct_stdized	14395
Trump16McCain08Shift_stdized	12064
perc_urban_stdized	5475
mean_temperature_stdized	-5140
median_age_stdized	-3609
share_in_apts_pct_stdized	502

Table 10: Lasso coefficients for covid deaths since 3/1/2020

Variable	**Lasso coef**
gini_inequality_stdized	0.0356
stringency_index_stdized	-0.0158
relative_humidity_stdized	-0.0097
hispanic_latino_pct_stdized	0.0089
nursing_resid_per_pop_stdized	0.0066
share_in_apts_pct_stdized	0.0062
afr_amer_pct_stdized	0.0061
under_18_pct_stdized	0.0054
population_per_sq_mi_stdized	0.0053
seldomornever_church_pct_stdized	-0.0034
mex_amer_pct_stdized	0.0022

Table 11: Lasso coefficients for excess death pct since 3/1/2020

Variable	Lasso coef
stringency_index_stdized	-39371
relative_humidity_stdized	-32552
nursing_resid_per_pop_stdized	30815
Trump16McCain08Shift_stdized	29510
hispanic_latino_pct_stdized	24536
population_per_sq_mi_stdized	23324
poverty_rate_stdized	22469
res_per_household_stdized	-17532
diabetes_pct_stdized	14640
seldomornever_church_pct_stdized	-13698
median_age_stdized	-12782
gini_inequality_stdized	2278

Table 12: Lasso coefficients for covid deaths since 10/1/2020

Variable	Lasso coef
mex_amer_pct_stdized	0.0338
stringency_index_stdized	-0.0273
gini_inequality_stdized	0.0232
relative_humidity_stdized	-0.0147
nursing_resid_per_pop_stdized	0.0147
seldomornever_church_pct_stdized	-0.0110
Trump16McCain08Shift_stdized	0.0110
obesity_rate_stdized	0.0057
under_18_pct_stdized	0.0028
median_age_stdized	-0.0025

Table 13: Lasso coefficients for excess death pct since 10/1/2020

5 Discussion

The author's expertise is in machine learning and data science, rather than epidemiology, virology, aerosol science, sociology or political science. The statistical associations are offered up in this paper as a starting point for further study by experts in these various subfields. With that said, some

cautious and tentative speculation for the assocations are presented in this section.

The consistent association of lockdown stringency with fewer deaths is certainly at least suggestive of a causal effect. The association may not necessarily prove the causal effect specifically of mandates, rules and regulations, though. One possible confounder is that states where it is politically palatable to enforce lockdowns are states where the population would have been more inclined to follow guidelines such as social distancing voluntarily even in the absence of mandates and rules. The evidence for the efficacy of social distancing and related measures (whether voluntary or enforced), however, seems strong. Perhaps there is some confounder which would explain the association of stringency with fewer deaths while undermining the case for the efficacy of public health guidelines, but no such confounder is apparent to the author at this time.

The causal pathway from higher humidity to reduced viral transmission seems clear based on research such as [22].

The assocation between inequality and covid deaths is open to many interpretations and it is possible that the relationship is not causal. With that said, many plausible explanations exist. Inequality may be associated with an increase in less-affluent workers providing various in-person services for more affluent residents (taxis and Ubers, restaurant waitstaff, house cleaners, hairdressers, etc). Inequality may also be associated with crowded living conditions in poor quality, poorly ventilated residencies. Inequality may also be associated with longer commute times either via public transportation or carpool, because high housing costs could lead to less affluent workers living farther away from their jobs. High-inequality state governments may also simply be less responsive to the needs and well-being of their less-affluent residents.

The reasons that Donald Trump gained vote share relative to previous Republican candidates are much studied and perhaps complex from a socieconomic point of view. There are many possible causal pathways which could explain the statistical relationship between Trump vote gain and higher covid deaths. Socieconomic conditions which led to Trump's appeal may have also caused higher covid deaths. With that said, the rhetoric and messaging of President Trump and his administration regarding social distancing, masks, etc also seems like a plausible causal pathway here.

6 Conclusions

There are many possible future directions for this research. The Oxford stringency index can be decomposed into several more specific mitigation measures. Statistical associations (or lack thereof) for these more specific measures are worthy of further study. The time dynamics of lockdown measures could also be studied more. This paper simply looks at the statistical association of average lockdown stringency with deaths over the entire course of the pandemic (or after October 1). Correlations between stringency at one point in time and covid deaths after a time lag (e.g. one month) merit some analysis.

The tools of modern causal inference, beyond just multiple regressions, such as directed acyclic graphs, g-computation and related techniques may also be useful here.

This paper may also serve as a warning against drawing conclusions about the relevance (or lack thereof) of variables based on simple pairwise correlations with a variable of interest. Lockdown stringency and humidity both show minimal correlation with covid deaths on their own, yet show up as extremely significant, in a very robust way, once a few other variables are included as controls.

References

[1] Gini coefficient. *Wikipedia*. https://en.wikipedia.org/wiki/Gini_coefficient.

[2] Household income: 2019.

[3] Lasso (statistics). *Wikipedia*. https://en.wikipedia.org/wiki/Lasso_(statistics).

[4] Oxford stringency index calculation. *Blavatnik School of Government*. Thomas, Webster, Petherick, Phillips, and Kira, www.bsg.ox.ac.uk/sites/default/files/Calculation%20and%20presentation%20of%20the%20St

[5] California mandated masks. florida opened its restaurants. did any of it matter? *Vox.com*, 2021. https://www.vox.com/coronavirus-covid19/22456544/covid-19-mask-mandates-lockdown-debate-evidence.

[6] Cases and death counts by place. *Washington Post*, 2021. https://www.washingtonpost.com/graphics/2020/national/coronavirus-us-cases-deaths/.

[7] J. L. Doti. Examining the impact of socioeconomic variables on covid-19 death rates at the state level. *J Bioecon*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7980794/.

[8] J. L. Doti. There's a better way than california's covid approach. *Orange County Register*. https://www.ocregister.com/2021/03/15/theres-a-better-way-than-californias-covid-approach/.

[9] M. et al. County-level predictors of coronavirus disease. *Clinical Infectious Diseases*, 2020. https://doi.org/10.1093/cid/ciaa1729.

[10] Y. Gu. Covid-19 state-level data. *Github*, 2021. https://github.com/youyanggu/covid19-datasets/blob/main/us_states_misc_stats.csv.

[11] Y. Gu. Thread. *Twitter*, 2021. https://twitter.com/youyanggu/status/1397230156301930497.

[12] Y. Gu. Tweet. *Twitter*, 2021. https://twitter.com/youyanggu/status/1397230177307004933.

[13] Y. Gu. Tweet. *Twitter*, 2021. https://twitter.com/youyanggu/status/1397247323189784585.

[14] L. C. Marr. Thread. *Twitter*. https://twitter.com/linseymarr/status/1343318611218345987.

[15] H. I. Moriyama. Seasonality of respiratory viral infections. *Annual Review of Virology*, 2020. https://www.annualreviews.org/doi/10.1146/annurev-virology-012420-022445.

[16] C. Shalizi. Lecture 26, variable selection. *Carnegie Mellon Statistics Course*. https://www.stat.cmu.edu/ cshalizi/mreg/15/lectures/26/lecture-26.pdf.

[17] J. Sill. U.s. state covid analysis. *Github*, 2021. https://github.com/jsill/usstatecovidanalysis.

[18] P. P. Thomas, Webster and Kira. Oxford covid-19 government response tracker. https://covidtracker.bsg.ox.ac.uk/.

[19] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x.

[20] D. J. Trump. Statement. -. https://www.donaldjtrump.com/news/statement-by-donald-j-trump-45th-president-of-the-united-states-of-america-06.12.21-01.

[21] O. S. K. Tsugawa. Association between state-level income inequality and covid-19 cases and mortality in the usa. *Journal of General Internal Medicine*, 2020. https://link.springer.com/article/10.1007/s11606-020-05971-3.

[22] L. C. M. Wan Yang. Dynamics of airborne influenza a viruses indoors and dependence on humidity. *PLOS ONE*, 2011. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021481.